Amazon Data Sciences

Recommendation Systems

- Julian Kahn
- Harmonyx Solutions LLC
- July 2, 2023

Contents and Agenda

- Historical context
- Business Problem Overview and Solution Approach
- User-User similarity based model
- Item-Item similarity based model
- Comparison of model performance
- Model Performance Summary
- Appendix

Historical Context

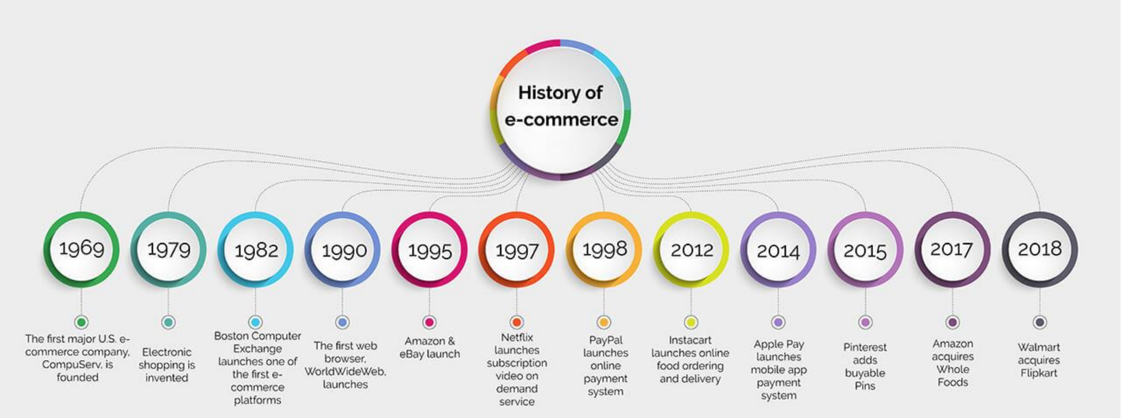
- **Early adoption and emergence:** E-commerce product ratings began to gain popularity in the late 1990s and early 2000s, as online shopping became more widespread. Websites like Amazon started allowing customers to rate and review products, providing valuable feedback and influencing purchase decisions. (Source: "The History of Online Product Reviews" - Trustpilot)

- **Influence on consumer behavior:** The introduction of product ratings had a significant impact on consumer behavior. Studies have shown that consumers heavily rely on product ratings and reviews when making purchase decisions online, considering them as a form of social proof and trust-building indicators. (Source: "The Impact of Online Reviews on Consumer Purchase Intention" - Journal of Retailing)

- **Importance for seller reputation:** Product ratings became an essential aspect of establishing and maintaining seller reputation in e-commerce. Positive ratings and reviews serve as a credibility signal for sellers, increasing customer trust and attracting new buyers. On the other hand, negative ratings can harm a seller's reputation and potentially affect sales. (Source: "Effects of eWOM and Product Ratings on Consumer Behavior: An Empirical Study" - International Journal of Information Management)

- **Development of rating systems:** Over time, e-commerce platforms and websites have refined their rating systems to improve the accuracy and relevance of product ratings. They introduced features such as verified purchase reviews, helpfulness voting, and detailed rating categories (e.g., quality, shipping, customer service), enhancing the overall rating experience for users. (Source: "Designing Trustworthy Online Rating Systems: The Role of Review Volume and Review Variance" - Journal of Management Information Systems)

- **Continuous evolution and challenges:** As e-commerce continues to evolve, new challenges have emerged in the realm of product ratings. These challenges include fake reviews, review manipulation, and biased ratings. E-commerce platforms have responded by implementing algorithms and moderation techniques to combat such issues and maintain the integrity of their rating systems. (Source: "Deceptive Opinion Spam Detection in Reviews" - ACM Transactions on Intelligent Systems and Technology)



Business Problem Overview

- As a Data Science Manager at Amazon, I have been assigned the business problem of creating a recommendation system – either user-user based or item-item based.

- OBJECTIVE:

  Utilize a collection of labelled data consisting of Amazon product reviews.

- GOAL:

  Derive valuable insights from the data and develop a recommendation system that effectively suggests products to online consumers.

Business Problem Overview contd.

- Information overload and **too many choices** create a **dilemma** for consumers, leading to **denial**.
- Recommender Systems provide **personalized** product recommendations to address this challenge.
- E-commerce giants like Amazon, Walmart, Target, and Etsy **invest** significant resources in developing **algorithmic techniques** for **personalized** recommendations.
- Amazon's recommendation system, known for its **accuracy**, **analyzes** and **predicts** customers' preferences to offer tailored product suggestions.
- Amazon utilizes **item-to-item** collaborative filtering, a baseline recommendation model that scales to large datasets and provides real-time, high-quality recommendations.

Etsy

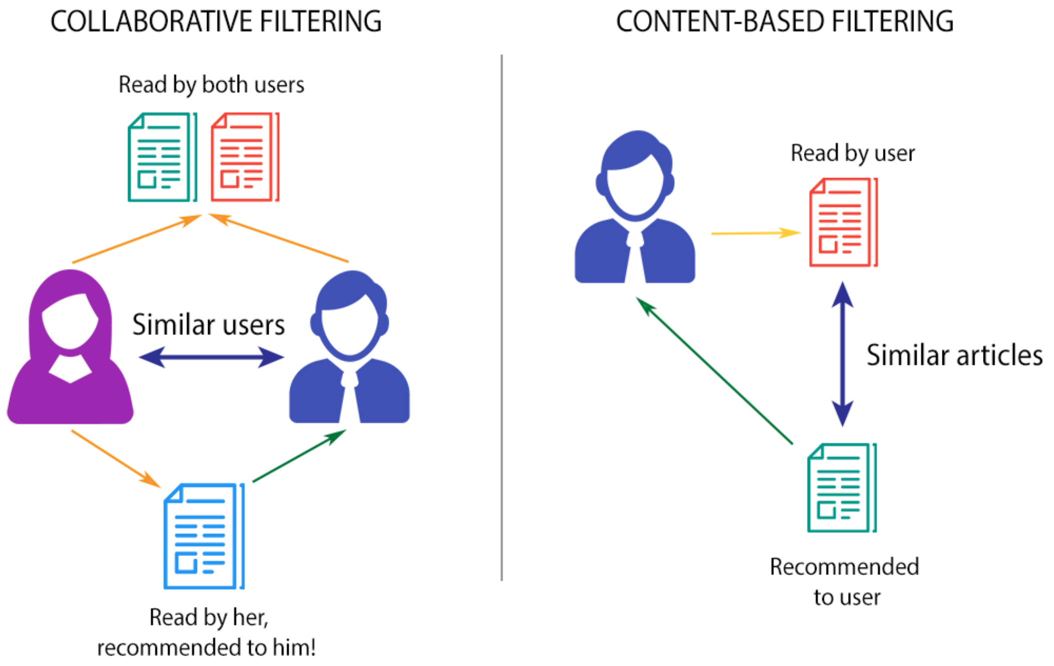
amazon.com

Walmart

Business Solution Overview

MODEL BUILDING - Types

- **Rank based recommendation system:** The simplest method of creating recommendation systems, where we assume that all customers have a similar preference. Unpersonalized recommendation system used when there isn't other data to go on.
- **Collaborative filtering-based recommendation system**
  - **User-user collaborative filtering** recommendation based on similar users (ie. Similar users have looked at this or purchased this)
  - **Item item collaborative filtering** recommendation based on previous purchases that have similarities between them
- **Singular value decomposition based collaborative filtering** matrix factorization
- **Content based recommendation system** (text based. This does not apply because we do not have text and can't apply natural language processing)



Business Solution Overview contd.

PERFORMANCE METRICS

- **RMSE (root mean square error)**

A popular metric used to evaluate the accuracy of a predictive model or the quality of predictions. It quantifies the average difference between the predicted values and the actual values in a dataset.

RMSE is calculated by taking the square root of the mean of the squared differences between the predicted and actual values. This ensures that the negative and positive errors do not cancel each other out, giving a more meaningful measure of the prediction error.

RMSE provides a measure of the model's performance in the same units as the target variable, making it easier to interpret. A lower RMSE indicates better predictive accuracy, as it signifies smaller average differences between predicted and actual values, while a higher RMSE suggests larger prediction errors.

- **Mean Absolute Error (MAE)**

Measures the average absolute difference between the predicted values and the actual values in a dataset.

Calculated by taking the mean of the absolute differences between the predicted and actual values. Unlike RMSE, MAE does not square the errors, which makes it more robust to outliers. MAE provides a straightforward measure of the average magnitude of prediction errors, regardless of their direction. A lower MAE indicates better predictive accuracy, as it signifies smaller average absolute differences between predicted and actual values.

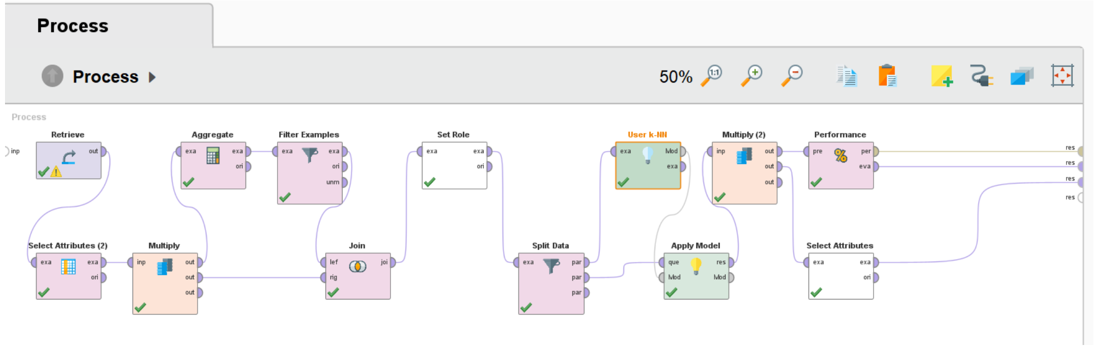
- **Normalized Absolute Mean (NAM/NMAE)**

A variant of Mean Absolute Error (MAE) that scales the error metric by dividing it by the range of the target variable, providing a normalized measure of the prediction accuracy.

- **Precision@k**

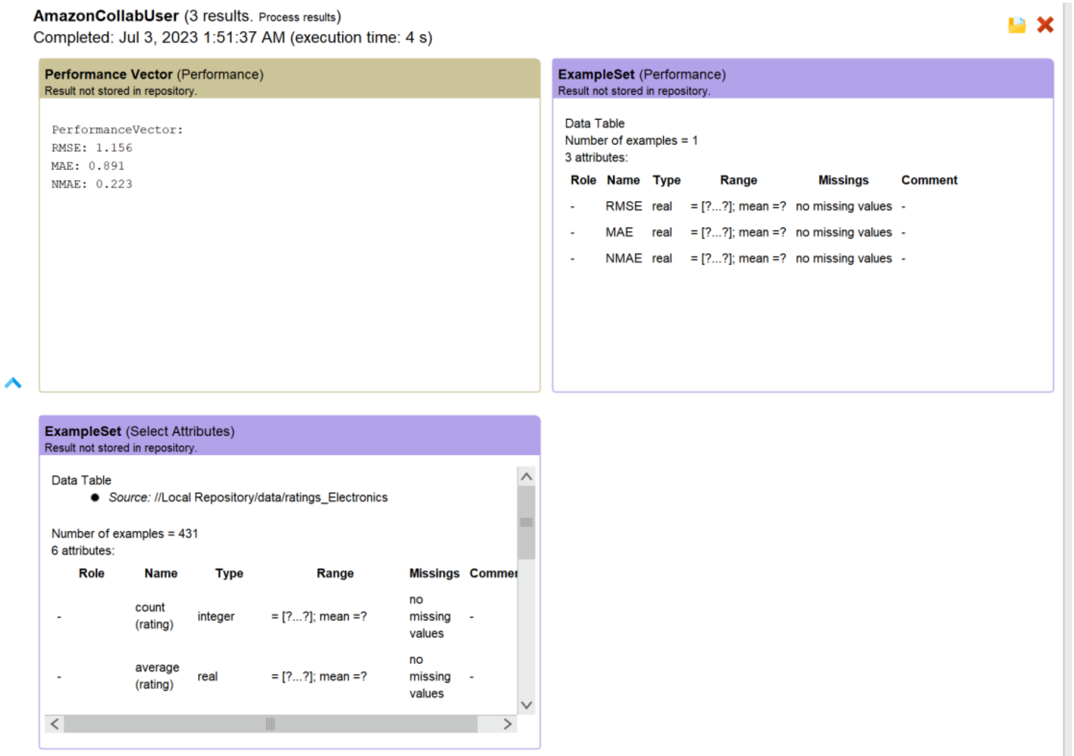
An evaluation metric used in information retrieval and recommendation systems to measure the proportion of relevant items or documents among the top k items recommended or retrieved. The higher the K, the better.

User-User Similarity Based Model



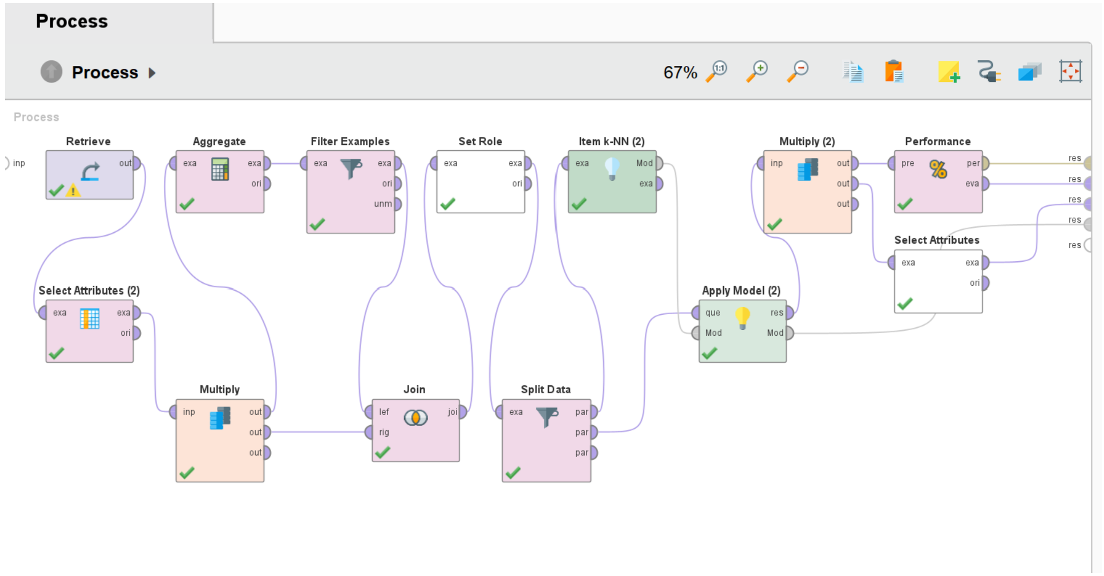
- Retrieve
- Select attributes
- Multiply -> Aggregate -> Filter Ex
- Join
- Set Role
- Split Data -> User K-NN
- Apply Model
- Multiply -> Performance
- Multiply -> Select Attributes

User-User Similarity Based Model



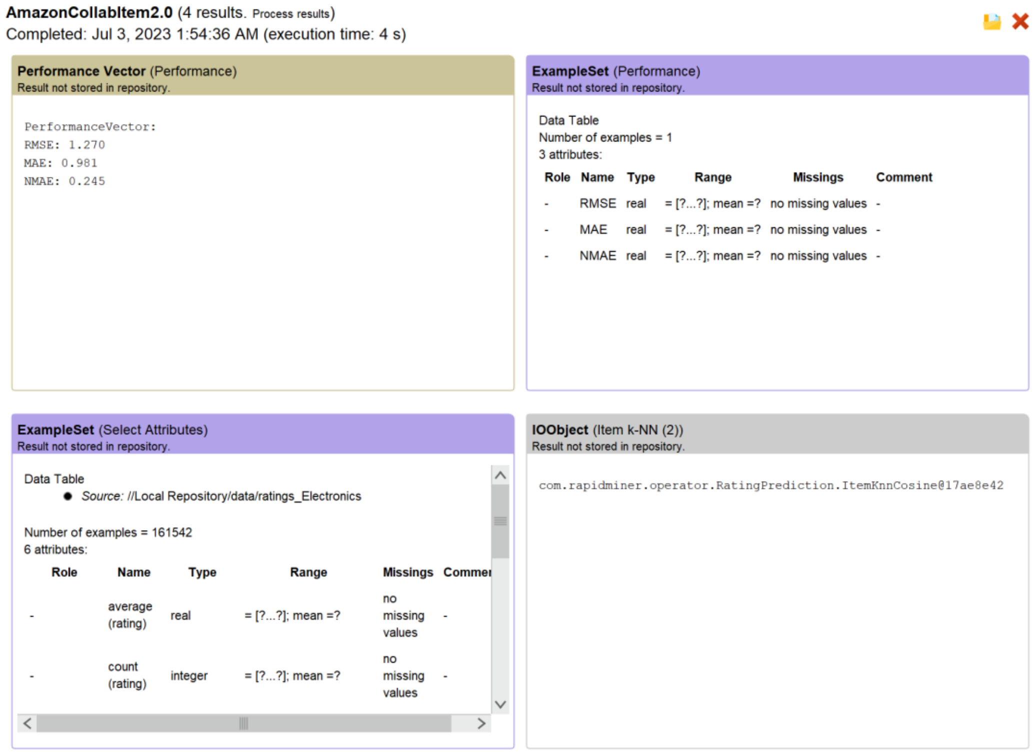
- RMSE: 1.156
- MAE: 0.891
- NMAE: 0.223

Item-Item Similarity Based Model



- Retrieve
- Select attributes
- Multiply -> Aggregate -> Filter Ex
- Join
- Set Role
- Split Data -> User K-NN
- Apply Model
- Multiply -> Performance
- Multiply -> Select Attributes

Item-Item Similarity Based Model



- RMSE: 1.270
- MAE: 0.981
- NMAE: 0.245

Comparison of Model Performance

- User-User
  - RMSE: 1.156
  - MAE: 0.891
  - NMAE: 0.223

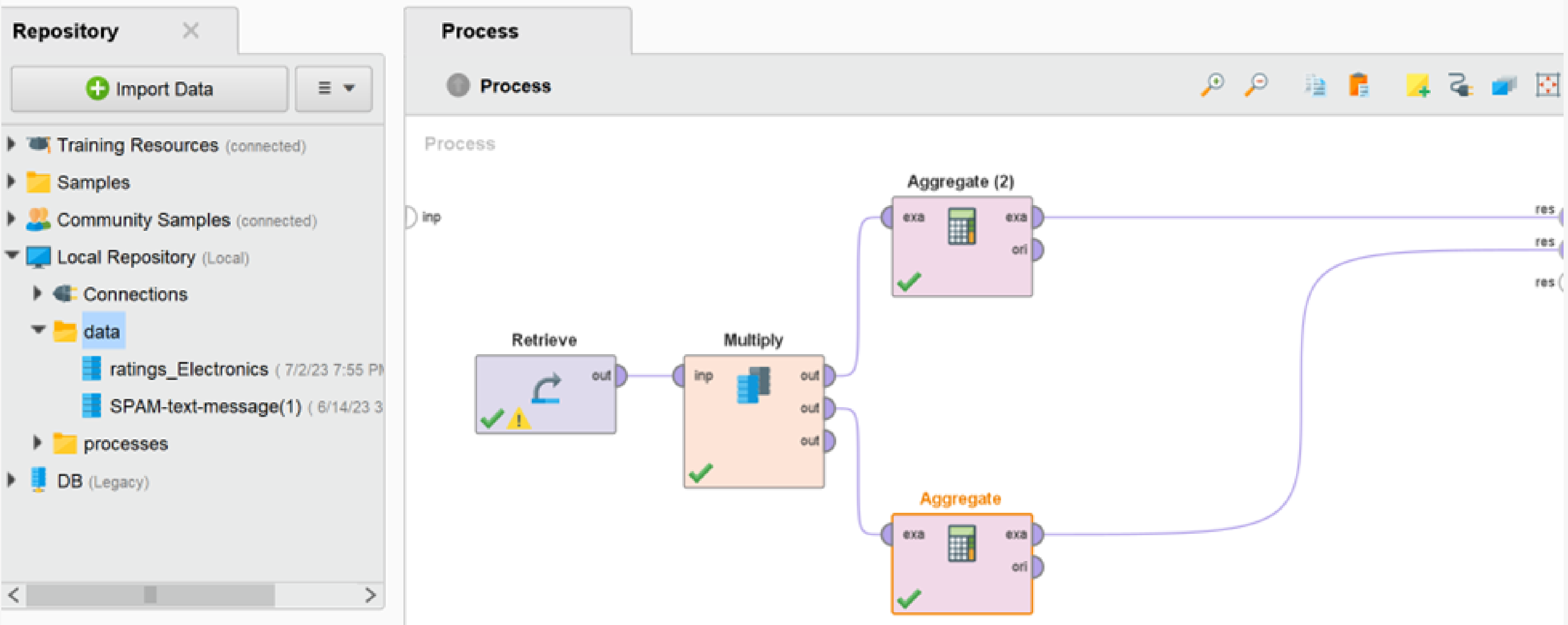
  In summary, user-user is more efficatiousthan item-item as a model as it shows lower values across the board, and therefore shows better predictive accuracy.

- Item-Item
  - RMSE: 1.270
  - MAE: 0.981
  - NMAE: 0.245

Appendix

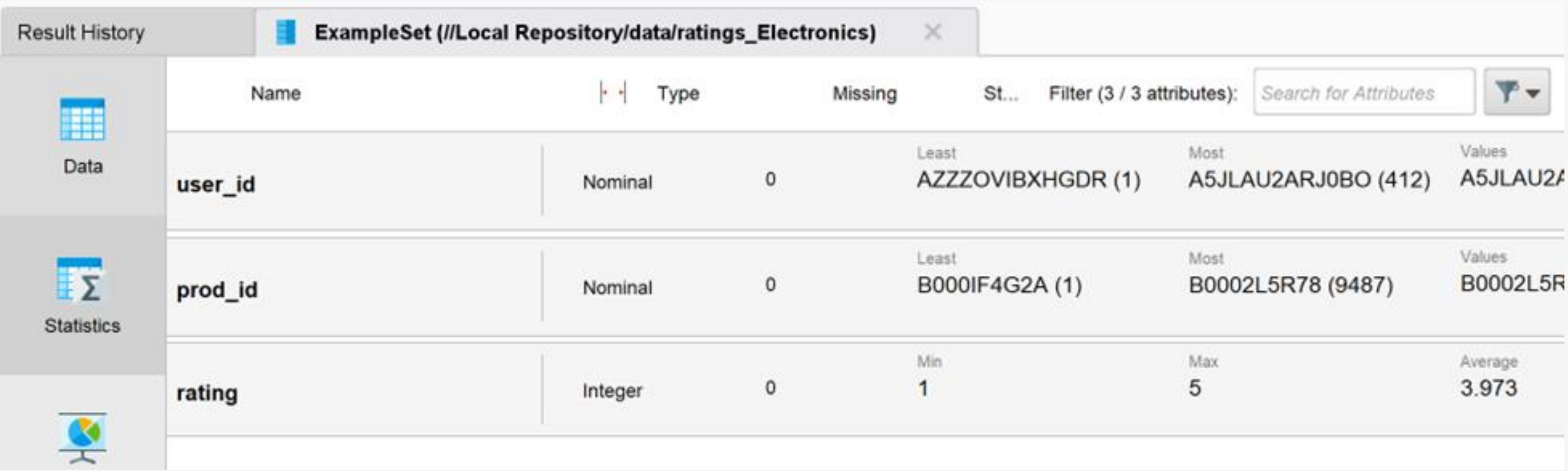
- Exploratory Data Analysis

- Statistics
- Visualization
- Ranked-based



Exploratory Data Analysis

Design



Statistics

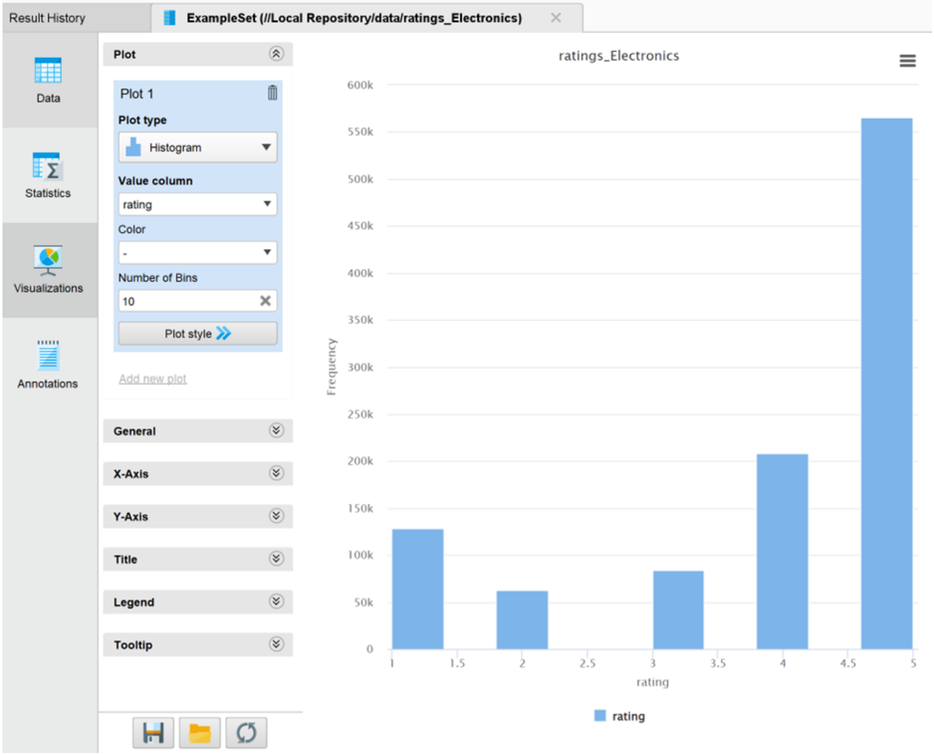
Statistics showed no missing columns ie. three rows of clean data (SKU, User, and Rating).

Statistics contd.

Summary

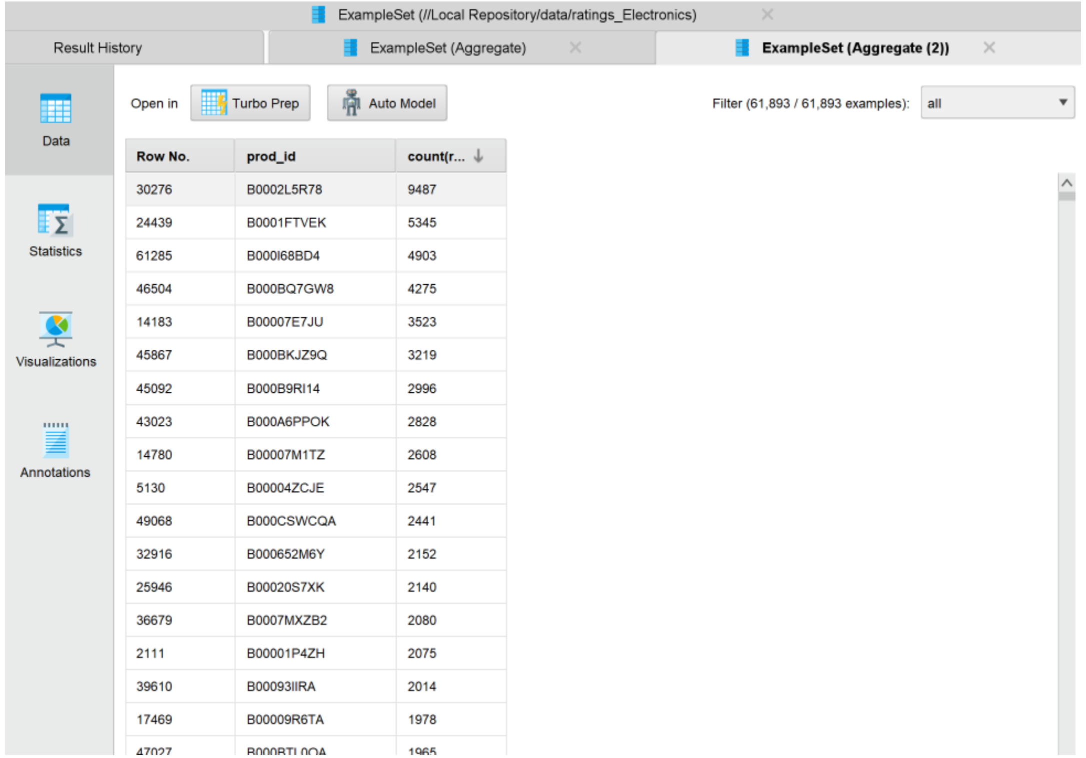
- No missing attributes and no missing entries
- Average user rating is 3.973. This indicates that products, when reviewed, are of a certain quality (or better) and, thus, receive favorable ratings.
- We are not reviewing text ratings, only ratings on a 1-5 basis, so we are not taking into account certain verbiage/context. We only have numerical ratings to make assertions from. This means that users are either not encouraged to provide written reviews to go along with their numerical review, are not motivated to provide a written review, or the project assignor has failed to provide written review data for analysis.

Exploratory Data Visualization



- We use histograms to: 1. Graphically summarize data 2. For strategic decision making purposes
- Good ratings from users, with a majority of ratings being 4 and 5 star reviews.
- Of the approximately 800k reviews of the 61893 product SKUs with reviews, 175k reviews are 1 or 2 star, so approximately 20% of reviews are negative in nature, 10% are neutral (three stars), and the remainder are generally positive, meaning 70% are 4 and 5 star reviews.

Exploratory Data Visualization contd.

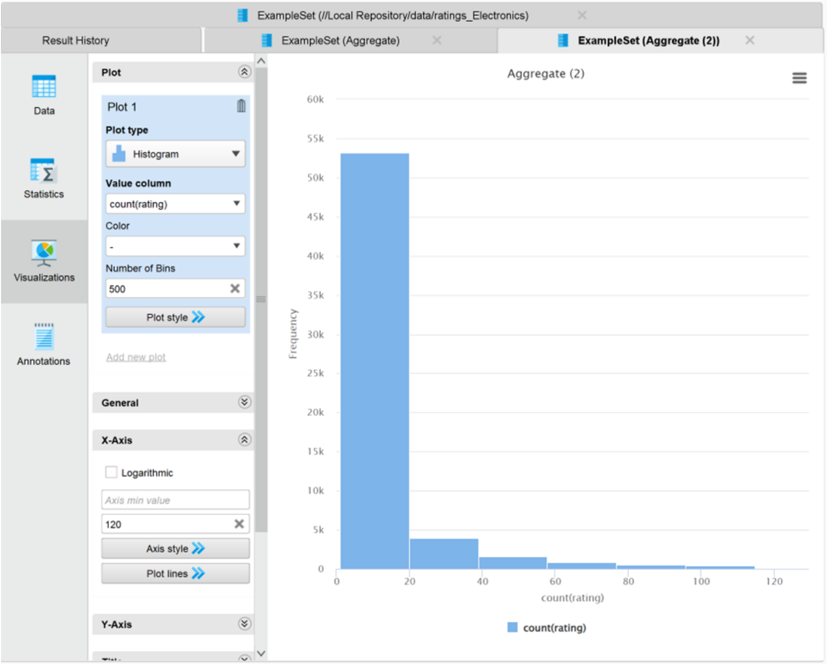


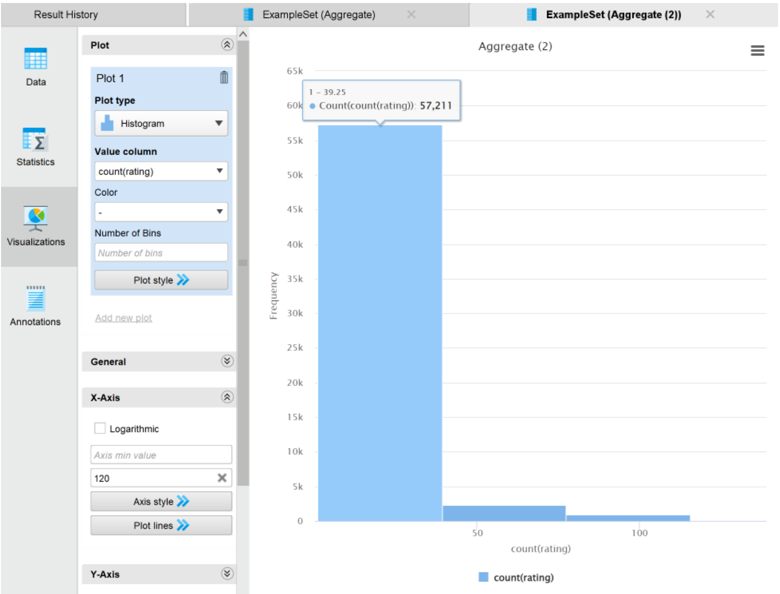
Row No.	prod_id	count(r...
30276	B0002L5R78	9487
24439	B0001FTVEK	5345
61285	B000I68BD4	4903
46504	B000BQ7GW8	4275
14183	B00007E7JU	3523
45867	B000BKJZ9Q	3219
45092	B000B9RI14	2996
43023	B000A6PPOK	2828
14780	B00007M1TZ	2608
5130	B00004ZCJE	2547
49068	B000CSWCQA	2441
32916	B000652M6Y	2152
25946	B00020S7XK	2140
36679	B0007MXZB2	2080
2111	B00001P4ZH	2075
39610	B00093IIRA	2014
17469	B00009R6TA	1978
47027	B000BTL0OA	1965

- The visual shows the SKUs with highest number of ratings
- There are 61893 ratings. This means 61893/786327 = **7.87%** of purchases have been reviewed (from a ML perspective, this could be considered low). Most products are thus left unreviewed adequately. BUT you can infer that very good or very bad products will elicit an emotional response and should be heeded.

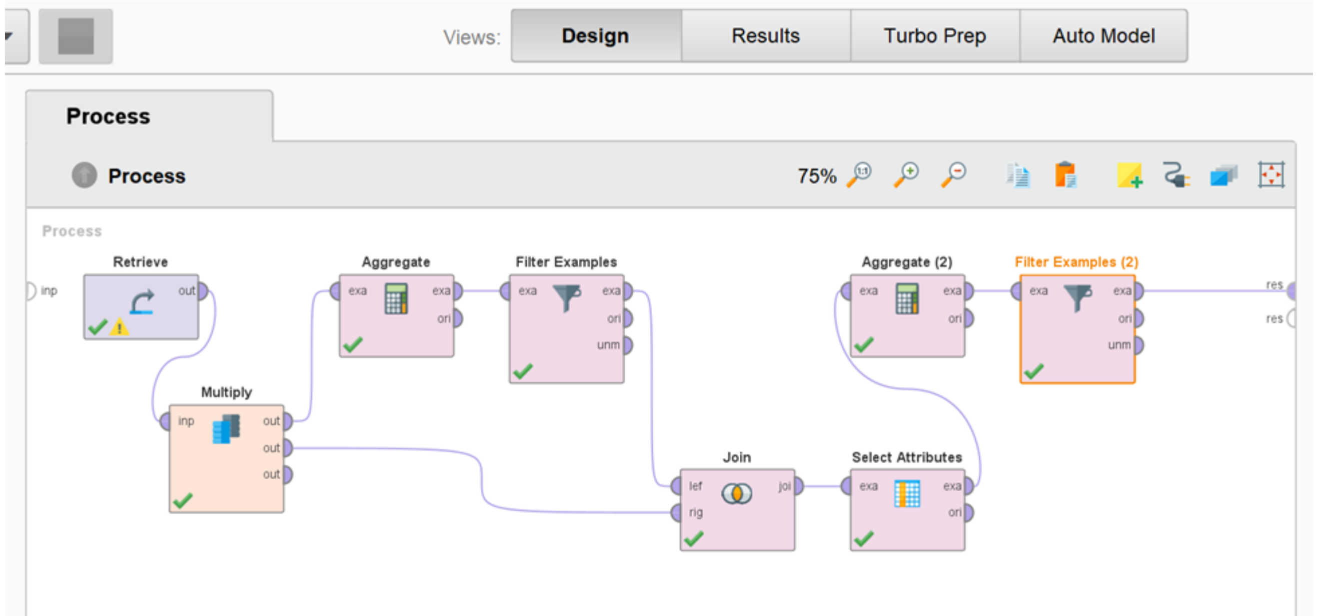
Exploratory Data Visualization contd.

- After increasing number of “bins” to 500, you can see more clearly the count rating spread.
- Customer reviews ranged from 0 to over 100





Ranked-based decision system



Model built showing reviewers with a minimum of 3 ratings for SKUs that have at least 50 reviews.

Users that have a minimum of 3 reviews for SKUs purchased might show a higher willingness/thoughtfulness to share their experience, thus being more trustworthy as reviewers. Of course, with no text data, this is still considered conjecture.

Ranked-based decision system

- We have now created a more refined dataset with 604 examples, vs the >800k examples we had to begin with. This is a significant *downselection* using customers with at least 3 reviews for products that, in aggregate, have 50 or more reviews.
- These can also be considered the more popular products that can be recommended to other users.

Result History | ExampleSet (Filter Examples (2)) | ExampleSet

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model

Row No.	prod_id	count(rating)	average...
117	B00006I53W	64	4.938
49	B000053HC5	79	4.911
222	B000144I2Q	53	4.906
371	B0007TOR08	51	4.902
379	B0007WK8LC	82	4.890
286	B0002GX0ZE	56	4.875
118	B00006I53X	123	4.862
591	B000I1X3W8	107	4.860
204	B0000BZL1P	383	4.859
50	B000053HH5	140	4.857
159	B000092TT0	158	4.854
47	B0000511U7	53	4.849
256	B00020M1U0	52	4.846
166	B000097O5F	58	4.845
385	B0007Y791C	69	4.841
310	B0002XQJFA	54	4.815
471	B000BYCKU8	134	4.813
558	B000FNFSPY	63	4.810
4	B00000J1V5	94	4.809
398	B0009GZAGO	73	4.808
378	B0007WK8KS	52	4.808

ExampleSet (604 examples, 0 special attributes, 3 regular attributes)